DOCUMENT RESUME

ED 325 504 TM 015 735

AUTHOR	Mullis, Ina V. S.; And Others
TITLE	The NAEP Guide: A Description of the Content and Methods of the 1990 and 1992 Assessments.
INSTITUTION	Educational Testing Service, Princeton, N.J.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NAEP-21-TR-01
PUB DATE	Apr 90
CONTRACT	RS89046001
NOTE	68p.
AVAILABLE FROM	National Assessment of Educational Progress (NAEP), Educational Testing Service, Rosedale Road, Princeton, NJ 08541.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Academic Achievement; Comparative Analysis; *Data Collection; *Educational Assessment; Elementary Secondary Education; *National Surveys; *Research Methodology; Sampling; State Surveys; Test Bias; Testing Problems; Test Interpretation
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

The National Assessment of Educational Progress (NAEP) is a survey of the educational achievement of American students and changes in their achievement across time. The NAEP has collected information for over 20 years to assist educators and policy makers. This guide intends to: document current NAEP methods for both state and national assessments; increase understanding of the philosophies and procedures of the NAEP; and demonstrate the consistency of the NAEP design as an indicator of academic achievement, including the 1990 and 1992 Trial State Assessment program. The guide is organized around 17 questions and answers covering: (1) the nature of the NAEP and identifying characteristics of the 1990 and 1992 assessments; (2) the NAEP's organization and management; (3) how the NAEP meets its goals; (4) trends and new assessments for 1990 and 1992; (5) the size of the NAEP sample; (6) how subject matter is determined; (7) the nature of the assessment questions; (8) bias against population groups in the NAEP; (9) contextual background data provided by the NAEP to help decision makers interpret the achievement results; (10) sampling methods; (11) how students are selected; (12) how cooperation of schools is ensured; (13) scoring problems; (14) analysis of results; (15) reporting of results; (16) the NAEP scales; and (17) reports for the 1990 and 1992 assessments. Fourteen publications are cited, which provide additional information about NAEP procedures and results. (SLD)

ED325504

The NAEP Guide

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Description of the Content and Methods of the 1990 and 1992 Assessments

Prepared by EDUCATIONAL TESTING SERVICE under a contract from THE NATIONAL CENTER FOR EDUCATION STATISTICS

Office of Educational Research and Improvement
U.S. Department of Education

THE NATION'S REPORT CARD

TM015735

The Nation's Report Card, the National Assessment of Educational Progress (NAEP), is funded by the U.S. Department of Education under a contract to Educational Testing Service. National Assessment is an education research project mandated by Congress to collect data over time on the performance of young Americans in various subject areas. It makes available information on assessment procedures to state and local education agencies.

This report, No. 21-TR-01, can be ordered from the National Assessment of Educational Progress at Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541.

The work upon which this publication is based was performed pursuant to Contract No: RS89046001 of the Office of Educational Research and Improvement.

Educational Testing Service is an equal opportunity/affirmative action employer.

Educational Testing Service, ETS, and ETS, are registered trademarks of Educational Testing Service.

The NAEP Guide

A Description of the Content and Methods of the 1990 and 1992 Assessments

Ina V.S. Mullis

in collaboration with

Albert E. Beaton, Jules M. Goodison, Eugene G. Johnson, Walter B. MacDonald, and Robert J. Mislevy.

John L. Barone, Ann Jungeblut, Stephen L. Koffler, Archie E. Lapointe, Donald Rock, and Rebecca Zwick also contributed as reviewers and their suggestions are much appreciated. Special thanks also go to reviewers from Westat, Inc., Keith F. Rust and Renee F. Slobasky. The editorial assistance provided by Lynn B. Jenkins and Peter Mann, the production assistance by Jan Askew and Kent Ashworth, and the word processing skills of Beverly A. Cisney are also gratefully acknowledged.

April 1990

The National Assessment of Educational Progress

Prepared by Educational Testing Service under a contract from The National Center for Education Statistics

• Office of Educational Research and Improvement •
U.S. Department of Education

TABLE OF CONTENTS

INTRODUCTION 5

QUESTION 1.
What, in brief, is the National Assessment of Educational Progress (NAEP)? Why are the 1990 and 1992 assessments special? 7

QUESTION 2.
How is NAEP currently organized and managed? 9

QUESTION 3.
How does NAEP meet its simultaneous--and conflicting--goals of both measuring trends in educational performance and providing information about student achievement on forward-thinking curricular goals? 12

QUESTION 4.
What trend and new assessments are being conducted in 1990 and 1992? 13

QUESTION 5.
How many schools and students are involved in the 1990 and 1992 national and state assessments? When are the data collected? 18

QUESTION 6.
Who decides what subject-matter content is measured by NAEP? How forward-looking are the 1990 and 1992 assessments? 22

QUESTION 7.
What are the assessment questions like? Do they keep pace with curricular advances? 24

QUESTION 8.
What efforts are made to ensure that the NAEP achievement measures are not biased against any population groups? 31

QUESTION 9.
What contextual background data does NAEP provide to help decision makers interpret the achievement results? 32

QUESTION 10.
Does NAEP use matrix sampling to reduce the burden for participating students? What is "Focused-BIB Spiraling" and what are the advantages of using it in NAEP? 37

QUESTION 11.
How are students selected for participation in NAEP? 40

QUESTION 12.
Who ensures the cooperation of sampled schools and administers the NAEP assessments and questionnaires? 42

QUESTION 13.
How does NAEP reliably score millions of open-ended responses without delaying the reports? How is the open-ended scoring merged with the computerized scoring of multiple-choice questions? 43

QUESTION 14.
How does NAEP analyze the assessment results? 46

QUESTION 15.
How can NAEP report results in a timely manner when procedures, and thus the computerized systems designed to implement those procedures, change from assessment to assessment? 53

QUESTION 16.
What are the NAEP scales? 54

QUESTION 17.
What types of reports will NAEP produce based on the 1990 and 1992 national and state assessments and when will they be available? 62

FURTHER READING 66

THE NAEP GUIDE

A Description of the Content and Methods of the 1990 and 1992 Assessments

INTRODUCTION

The National Assessment of Educational Progress (NAEP) is a survey of the educational achievement of American students, and changes in that achievement across time. Fashioned in 1969 as an educational indicator, NAEP has successfully collected information for over 20 years with the philosophy of providing accurate and useful information to educators and policy makers while placing as little data collection burden as possible on students.

As the nation's primary indicator of what schoolchildren know and can do, NAEP's utility and credibility are based on its ability to change and keep pace with current interests in education. NAEP makes conscientious attempts to reflect changes in curriculum and in educational objectives. These efforts to be sensitive to changing school environments necessitate changes in the assessment each time a curriculum subject is measured. NAEP also makes conscientious efforts to respond to changes in assessment technology, for example, incorporating increasingly refined IRT scaling methods into the data-analysis procedures and adopting innovations in performance testing, including the science and mathematics "hands-on" pilot, as well as special studies in mathematics problem-solving and estimation, oral reading, and portfolios of reading and writing samples.

Because NAEP is the assessment instrument for a diverse nation imbued with the spirit and tradition of freely voicing opinions, it is quite appropriate that the process of developing, conducting, analyzing, and reporting NAEP is implemented under the guidance of and with input from an ongoing series of meetings and reviews involving outside experts and governmental agencies. Inevitably, each discussion suggests useful revisions in some aspect of the system, and NAEP has been flexible enough to accommodate many of these changes and evolve from assessment to assessment.

If NAEP were based solely on inflexible computerized analysis and a reporting system that required no data-contingent decisions, very few people would find the results useful. However, every change ripples through the system, requiring research to implement new approaches in accurate and efficient ways. Because NAEP is central to our nation's evaluation of its condition and progress in education, it is important to devote significant amounts of time to quality control.

Currently, due to the President's Summit on Education and the resultant Education Goals, NAEP is playing an increasingly visible role in measuring student achievement. Also, as NAEP enters the 1990s, it will be providing state level results for the first time. Because of the need for assessment data, there is increased pressure to reduce the reporting time for NAEP results to make the information more readily available to education policy makers. With each assessment, significant efficiencies have been incorporated that have reduced the amount of time from final data collection to the issuance of attractive, reliable, easily understood reports. Being both responsive and responsible, however, takes time.

The purpose of this guide is to document the current NAEP methods for both the national and state assessments, to increase understanding of the philosophies and procedures underlying NAEP, and to highlight the consistency of NAEP's design with its role as an educational achievement indicator, including the 1990 and 1992 Trial State Assessment program.

#1. Question: What, in brief, is the National Assessment of Educational Progress (NAEP)? Why are the 1990 and 1992 assessments special?

Answer: *The National Assessment of Educational Progress (NAEP) is a primary indicator of the level of our students' academic achievement.*

Since 1969, NAEP has been assessing what American students know and can do in a variety of curriculum areas and plotting their progress across time. To provide context for the achievement results, NAEP also collects demographic, curricular, and instructional background information from students, teachers, and school administrators. Also known as The Nation's Report Card, NAEP publishes these results in a series of widely disseminated reports.

In 1988, Congress added a new dimension to NAEP by authorizing, on a trial basis, voluntary participation in state-level assessments in 1990 and 1992. Designed to provide results comparable to the nation and other participating states, the trial state assessments include eighth-grade mathematics as well as fourth-grade mathematics and reading in 1992. With the President's Summit on Education, the resultant Education Goals, and the addition of the 1990 and 1992 trial state assessment program, NAEP is playing an increasingly visible role in measuring student achievement.

From its inception, NAEP has assessed the achievement of national samples of 9-, 13-, and 17-year-old students in public and private schools. In 1983, it expanded the samples so that grade-level results could also be reported. In 1990 and 1992, NAEP also will assess fourth, eighth, and twelfth graders.

The assessments, conducted annually until the 1979-80 school year and biennially since then, have included periodic measures of student performance in reading, mathematics, science, writing, U.S. history, civics, geography, and other subject areas. In 1990 and 1992, NAEP will assess reading, mathematics, science and writing. NAEP assessed U.S. history, civics, and geography as well as reading and writing in 1988.

At the historic education summit in Charlottesville, in the fall of 1989, the President and the Governors declared the "the time has come, for the first time in U.S. history, to establish clear national performance goals, goals that will make us internationally competitive." As part of those national education goals, it is stated that:

> "By the year 2000, American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy."

Objectives elaborating on this goal include increasing the academic performance of elementary and secondary students significantly in every quartile, the distribution of minority students to more closely reflect the student population as a whole, and the percentage of students who demonstrate the ability to reason, solve problems, apply knowledge, and write and communicate effectively.

NAEP augments such information on achievement, by asking students about their backgrounds, the courses they have taken, and their classroom activities (how they are taught). Their teachers are also asked to provide information about their backgrounds and training, the content emphases of the students' courses, the instructional approaches used, and the availability of resources. In addition, students' principals are asked about school staffing and resources as well as about policies related to teaching, learning, and curriculum. These data are related to student performance levels in an effort to inform policy, reinforce research evidence about factors that relate to school achievement, and help improve educational outcomes.

Forty jurisdictions--37 states, the District of Columbia, and two territories--participated in the 1990 trial state assessment program in eighth-grade mathematics and it is anticipated that participation in the 1992 program, which also includes fourth-grade reading and mathematics, will be at least at a similar level. Participating states may compare their results to the national statistics and, if they wish, to each other.

As the nation's only ongoing program to monitor academic achievement at the elementary, middle-school, and high-school levels, NAEP reports performance trends across time for different demographic subgroups and for different parts of the country. These data, as well as the forthcoming results for states, are collected and disseminated to help policy makers evaluate and improve the quality and equity of American education. NAEP data are cited regularly in major reports on education as well as in professional journals, newspapers, and magazines.

#2. Question: How is NAEP currently organized and managed?

Answer: *NAEP is a congressionally mandated project of the National Center for Education Statistics (NCES), U.S. Department of Education. NCES conducts NAEP through a series of grants and contracts, designed to fulfill the requirements of the NAEP legislation and be responsive to substantive and policy recommendations provided by the National Assessment Governing Board (NAGB). Educational Testing Service (ETS) was awarded the NAEP operations contract for the 1990 and 1992 assessments, including the trial state assessment component. However, the states are also a partner in this effort, because they are responsible for state data collection and because state representatives review the materials and procedures used in the trial state assessment program. In addition, NAEP is responsive to advice obtained from a broad variety of experts and interested parties concerned with improving assessment and education in America.*

More specifically, the responsibility for particular activities is specified by law (P.L. 100-297, enacted in 1988).

- The Commissioner of Education Statistics is responsible for carrying out the NAEP project through competitive awards to qualified organizations. NCES awarded the operational contract for the conduct of the 1990 and 1992 assessments to Educational Testing Service (ETS) and its subcontractors: Westat, Inc., which is responsible for sampling and data collection, and National Computer Systems (NCS), which is responsible for

printing, open-ended scoring, and scanning. In addition to coordinating operational activities, ETS is responsible for developing the assessment questions according to specifications provided to NCES by NAGB, analyzing the results, and working with NCES staff to prepare the reports on student achievement in the various subject areas assessed. The legislation authorizing the trial state assessments requires that they be evaluated to determine "the feasibility and validity of [state] assessments and the fairness and accuracy of the data they produce." A national commission of measurement experts, cochaired by Robert Linn and Robert Glaser, is conducting this evaluation under a grant from NCES. The commission is supported by staff from the American Institute of Research (AIR) and the National Academy of Education (NAE). Additional awards will be made for other NAEP activities, including conducting validity studies and reporting the results of additional in-depth analyses of NAEP data across subject areas.

- The National Assessment Governing Board (NAGB) formulates policy guidelines for NAEP. NAGB's composition is specified by law, and its 24 members include teachers, curriculum specialists, state legislators, governors, measurement experts, chief state school officers, state and local school board members, school superintendents, principals, and representatives from business and the general public. More specifically, NAGB is responsible for selecting subject areas to be assessed, in addition to those specified by Congress; identifying appropriate achievement goals for each age and grade; developing assessment objectives; developing guidelines and standards for data analysis and for reporting and disseminating results; developing standards and procedures for interstate, regional, and national comparisons; improving the form and use of the National Assessment; and ensuring that all items selected for use in the National Assessment are free from racial, cultural, gender or regional bias.

- States that participate in the trial state assessments are responsible "in cash or in kind" for test administration within their borders, including providing qualified personnel to accomplish this task, securing school participation, making sure that students in the sample attend the scheduled assessment sessions.

In addition to NCES, its grantees and contractors, NAGB, and state education agency personnel, NAEP solicits advice from numerous committees of experts as well as through public hearings, forums, and widespread mail reviews. For example, in developing objectives and specifications for the 1992 assessment, NAGB held public hearings and solicited written comment from experts and interested parties. The consensus development process was managed through a NAGB contract with the Council of Chief State School Officers (CCSSO) and its subcontractor, AIR. The CCSSO worked with a 16-member Steering Committee representing a variety of professional organizations, and a 16-member Planning Committee of reading and assessment experts, and solicited widespread review from state education agencies and reading educators. In the area of goal setting, NAGB held a public forum and obtained advice from a number of consultants. In addition, NAGB commissioned committees to guide policy on the 1992 writing objectives, state-by-state reporting, and analysis variables; it also commissioned papers on measuring students' opportunity to learn and on the NAEP timelines.

NCES and its contractors also devote considerable energy to obtaining advice from external sources. For example, ETS manages the NETWORK of participants to solicit advice and share information about the trial state assessments. Comprising state testing directors and their NAEP assessment coordinators as well as staff from NCES, NAGB, CCSSO, ETS, Westat, NCS, and AIR, the NETWORK meets quarterly to discuss the plans and progress of the trial program. The NETWORK News and the State of Education bulletins also keep interested parties informed.

To help develop the assessment items, ETS relies on external subject matter specialists, state curriculum coordinators, and testing experts. Each subject-area assessment has its committee of experts in the field, drawn from across the country. In addition, a separate committee considers the focus, validity, appropriateness, and utility of the background questions. ETS also regularly convenes the Design and Analysis Committee for NCES. This committee of prestigious statisticians, psychometricians, and measurement experts advises on technical issues spanning subject areas and assessment years.

In 1990, more than 5,000 people were involved in the development and implementation of both the national and state assessments. The NAEP data collection effort alone involved thousands of people, primarily local administrators in the states, who were trained by Westat to standardize the test administration, ensure that proper procedures were followed in collecting the test booklets and questionnaires, and observe test security. (For more information about NAEP data collection, please see #12.)

#3. **Question: How does NAEP meet its simultaneous--and conflicting--goals of both measuring trends in educational performance and providing information about student achievement on forward-thinking curricular goals?**

Answer: *NAEP uses several procedures to maintain the stability required for measuring trends, while still introducing innovations. To keep pace with developments in assessment methodology and research about learning in each subject area, NAEP updates substantial portions of the assessments with each successive administration. However, in some subject areas, NAEP also provides separately for links to the past and links to the future by conducting parallel assessments. For example, in mathematics and reading, to maintain trends established across 20 years, nationally representative samples of students are assessed using the methods of past assessments. In addition, to be fully responsive to recommendations from the consensus process about current thinking in the field, separate nationally representative samples of students also are assessed using newer, more innovative methods.*

To ensure that trend assessment results reflect changes in student performance and not changes in the test, many of the questions--and all of the procedures--are held constant from assessment to assessment. In general, for trend components of the assessment, NAEP uses an item replacement scheme that provides for updating approximately half the questions by including matched, but forward-looking replacements in the instrument for the next assessment.

However, this system does not provide for major changes in curriculum or in ways to conduct assessment. After 20 years of measuring trends using this system, the trial state assessment program provided an opportunity to "start over." Thus, beginning in 1990 for mathematics and in 1992 for reading and writing, NAEP is initiating whole new systems of measuring achievement for the future that are highly responsive to innovations in curriculum, instruction, and assessment methodology.

For several assessments in each subject area, however, both the old and new systems are being used. The prior set of content questions and procedures are still being used in their intact form to retain links to the past, while at the same time new trend lines are being established on curricular goals that are more suitable as we approach the 21st century--and on assessment methods more appropriate to those goals. When, at some time in the future, the new trend lines begin to emerge, assessments linked directly to the 1970s and 1980s can be discontinued.

#4. Question: What trend and new assessments are being conducted in 1990 and 1992?

Answer: *As summarized in the table below, in 1990 and 1992, long-term trend assessments linking results across NAEP's 21-year history will continue to be administered in mathematics, reading, science, and writing. Also, a newly designed mathematics assessment begins for the nation in 1990 that will be updated and carried forward to produce short-term trends at grades 4, 8, and 12 in 1992. In reading, the short-term trend assessment conducted in 1990 will be replaced in 1992 at all three grades with an elaborate new*

national assessment developed in conjunction with the trial state assessment program. The state assessments in mathematics at grades 4 and 8 and reading at grade 4 replicate most of the questions included in the newly developed national assessments. Additionally, for the nation, new procedures for assessing science will be used in 1990 and for writing in 1992.

Types of Assessments Included in 1990 and 1992

	Mathematics		Reading		Science		Writing	
	1990	1992	1990	1992	1990	1992	1990	1992
Long-term trend	X	X	X	X	X	X	X	X
New or short-term trend	X	X	X	X	X			X
State level	X	X		X				

Subject by subject, NAEP conducted the following assessments in 1990:

1990 Mathematics

- For the nation, a newly developed mathematics assessment was given at grades 4, 8, and 12 that included the use of scientific calculators and open-ended problem solving. Estimation and complex problem solving were assessed in a special study using audiotapes that paced students through the questions.

- For the 40 participating states and entities, the newly developed mathematics assessment was given at grade 8.

- A national long-term trend assessment of 9-, 13-, and 17-year-olds was administered, and the results will link five mathematics assessments conducted across the past 17 years (in 1973, 1978, 1982, 1986, and 1990).

1990 Reading

- A reading assessment, developed in 1988 and updated in 1990, included questions designed to measure reading as a process involving the construction and examination of meaning. Students at grades 4, 8, and 12 were asked multiple-choice as well as a few open-ended questions about literary and informational passages and about documents.

- A long-term trend assessment was administered at ages 9, 13, and 17 that links six reading assessments conducted across the past 19 years (1971, 1975, 1979, 1984, 1988, and 1990).

1990 Science

- A newly developed science assessment at grades 4, 8, and 12 included two types of open-ended questions--asking students to write brief responses demonstrating their ability to conduct scientific inquiry and to draw illustrations indicating their grasp of scientific events.

- The long-term trend assessment for 9-, 13-, and 17-year-olds was given to provide results spanning a 21-year period and six science assessments (in 1969, 1973, 1977, 1982, 1986, and 1990).

1990 Writing

- A national trend assessment was administered to fourth-, eighth-, and eleventh-grade students that linked directly to 1984 and 1988. Based entirely on writing performance (rather than multiple-choice questions), the assessment includes a variety of informative, persuasive, and narrative prompts, enabling NAEP to measure performance on individual tasks and on a scale across tasks. Results will include trends in students' ability to

accomplish a particular purpose in writing, their overall fluency, and the incidence (or prevalence) of grammatical and mechanical errors in their writing.

- A pilot portfolio study was conducted in conjunction with the trend assessment at grades 4 and 8. In this study, "The Nation's Writing Portfolio," students and their teachers were asked for examples of the students' best writing--in particular, writing that incorporated process strategies that are difficult to implement in a regular assessment (e.g., getting peer and teacher review, using external resources, and working on a paper over several weeks).

NAEP plans to conduct the following assessments in 1992:

1992 Mathematics

- For the nation, NAEP will administer a carefully updated version of the assessment newly developed for 1990 including the special study materials. Thus, the 1992 assessment will provide short-term trend information at grades 4, 8, and 12.

- For participating states and other entities (i.e., the District of Columbia and territories), NAEP will conduct comparable short-term trend assessments for eighth graders and assessments comparable to the nation and across states for fourth graders.

- The long-term trend assessment will be administered to extend the nearly 20 years of trend data for 9-, 13- and 17-year-olds.

1992 Reading

- Building on the experiences of the 1988 and 1990 assessments, an entirely new 1992 reading assessment is being designed for fourth, eighth, and twelfth graders. This assessment will require many more open-ended responses and contain longer, more natural looking, passages, including literary and informational texts as well as documents. Enough questions will be asked to permit reporting separate results for the three types of materials. At grade 4, the national assessment will include an oral reading and portfolio component.

- Fourth-graders in each state or other jurisdictions participating in the voluntary state program will be assessed on the newly developed reading assessment.

- The long-term trend assessment will provide updated trend results for 9-, 13-, and 17-year-olds.

1992 Science

- National long-term trend results for 9-, 13-, and 17-year-olds will be updated to include results for 1992.

1992 Writing

- A new writing assessment is being designed for the nation's fourth, eighth, and twelfth graders that will respond directly to the current instructional emphasis on the writing process. Based on 25- and 50-minute prompts, the assessment will ask students to plan and revise their writing, give them guidance as to how they will be evaluated, and judge the results accordingly.

- The writing portfolio assessment will be continued.

- The trend assessment for fourth, eighth, and eleventh graders will be readministered.

(Additional details about these assessments are provided in #5 and #6.)

#5. Question: How many schools and students are involved in the 1990 and 1992 national and state assessments? When are the data collected?

Answer: *Across grades and age levels, trend as well as newly developed national assessments and their comparable state-level assessments, and the various subject areas, each assessment involves many distinct systematic samples of students, thousands of schools, hundreds of thousands of individual students, and millions of written responses to open-ended questions. In 1992 alone, there will be 419,000 students comprising about 200 samples at 12,000 schools, and they will generate more than 7 million open-ended responses.*

Although trend assessments linked to previous procedures are conducted throughout the school year according to schedules used since the inception of NAEP, the bulk of the testing in 1990 was accomplished in January through May, including the state assessments which were conducted in February. In 1992, the state assessment schedule remains the same, but most of the national assessments will be conducted in January through March.

The following tables give detailed information on the sample sizes and data collection schedules for the various components of the 1990 and 1992 assessments.

1990 Newly Developed and Short-term Trend Assessments

Age/Grade	Subject*	Type of Session**	Booklets***	Students	Open-ended Responses	Schools	Data Collection
9/4	M	Spiral	7	8,000	115,000		1/8/90 to
	M	Tape	1	3,000	47,000	560	5/18/90
	R/S	Spiral	17	19,000	95,000		
13/8	M	Spiral	7	8,000	144,000	420	1/8/90 to
	M	Tape	1	3,000	25,000		5/18/90
	R/S	Spiral	14	16,000	116,000		
--/8 State (40 participated)	M	Spiral	7	100,000	1,835,000	3,900	2/5/90 to 3/9/90
17/12	M	Spiral	7	8,000	144,000		1/8/90 to
	M	Tape	1	3,000	47,000	300	5/18/90
	R/S	Spiral	14	16,000	120,000		

1990 Long-term Trend Assessments

Age/Grade	Subject*	Type of Session**	Booklets***	Students	Open-ended Responses	Schools	Data Collection
9/4	R/W	Spiral	6	5,200	38,000		1/8/90 to
	M/S	Tape	5	10,000	80,000	319	3/16/90
13/8	R/W	Spiral	6	5,200	49,000		10/9/89 to
	M/S	Tape	5	10,000	60,000	282	12/15/89
17/11	R/W	Spiral	6	5,200	51,000		3/19/90 to
	M/S	Spiral	6	7,800	80,000		
	M/S	Tape	4	8,000	72,000	310	5/18/90

* M = Mathematics, R = Reading, S = Science, W = Writing

** In "spiral" assessments, students respond to different assessment booklets that are evenly distributed across students. In "tape" assessments, all students respond to the same booklet accompanied by an audiotape that paces them through the items.

*** Number of different assessment forms.

1992 Newly Developed and Short-term Trend Assessments

Age/Grade	Subject*	Type of Session**	Booklets***	Students	Open-ended Responses	Schools	Data Collection
9/4	M	Spiral	7	8,000	120,000		1/6/92 to
	M	Tape	1	3,000	50,000	560	3/31/92
	R/W	Spiral	36	16,000	155,000		
--/4 State	R	Spiral	18	120,000	950,000	6,000	2/3/92 to
(45 anticipated)	M	Spiral	7	120,000	2,400,000		3/6/92
13/8	M	Spiral	7	8,000	150,000	420	1/6/92 to
	M	Tape	1	3,000	28,000		3/31/92
	R/W	Spiral	40	27,000	210,000		
--/8 State	M	Spiral	7	120,000	2,400,000	4,500	2/3/92 to
(45 anticipated)							3/6/92
17/12	M	Spiral	7	8,000	150,000		1/6/92 to
	M	Tape	1	3,000	50,000	300	3/31/92
	R/W	Spiral	41	30,000	233,000		

1992 Long-term Trend Assessments

Age/Grade	Subject*	Type of Session**	Booklets***	Students	Open-ended Responses	Schools	Data Collection
9/4	R/W	Spiral	6	5,200	38,000		1/6/92 to
	M/S	Tape	5	10,000	80,000	325	3/13/92
13/8	R/W	Spiral	6	5,200	49,000		10/7/91
	M/S	Tape	5	10,000	60,000	290	12/13/91
17/11	R/W	Spiral	6	5,200	51,000		3/16/92
	M/S	Tape	4	8,000	72,000	320	5/15/92

* M = Mathematics, R = Reading, S = Science, W = Writing

** In "spiral" assessments, students respond to different assessment booklets that are evenly distributed across students. In "tape" assessments, all students respond to the same booklet accompanied by an audiotape that paces them through the items.

*** Number of different assessment forms.

As shown, different kinds of assessments must be conducted at different times. The trend assessments have to be conducted along the same timelines as in all previous assessments; otherwise, the results will not be truly comparable. Thus, for these assessments, 13-year-olds are assessed in the fall, 9-year-olds in the winter, and 17-year-olds in the spring.

To improve NAEP's ability to compare student performance across age and grade levels, schedules for newly developed assessments have been set to permit simultaneous surveys at all three ages and grades. For example, in 1990, at all three age/grade levels, NAEP identified two equivalent samples of students and assessed each of them in all three subjects (mathematics, reading, and science). For convenience in coordinating the data collection activities, in each grade, half the students were assessed in the winter on the same schedule as the age 9 trend assessments and half the students were assessed in the spring on the same schedule as the age 17 trend assessments.

For the trial state assessments, however, data were collected in February, because a poll of state testing directors showed that they preferred that timing. To spread the data collection effort evenly within and across states, each state assessed students in one-fourth of its schools every week throughout the month.

In 1990, the gap in completing the state and national assessments meant the national figures necessary for comparative purposes trailed the availability of the state results by a considerable margin, thereby delaying state assessment reports. In 1992, the period for national data collection will be shortened by conducting the assessments in the winter, cutting two months off the timetable for reporting state results. By conducting future mathematics assessments from January through March, NAEP will be able to maintain trends with the 1990 mathematics assessment by comparing results to those obtained from the first half sample in that assessment. The 1992 writing and reading assessments are wholly new and, thus, also can be conducted during the winter time period.

#6. **Question: Who decides what subject-matter content is measured by NAEP? How forward-looking are the 1990 and 1992 assessments?**

Answer: *The subject-area objectives for each NAEP assessment are determined through a legislatively mandated consensus process managed by the National Assessment Governing Board (NAGB). These objectives typically take the form of frameworks or matrices delineating the important content and process areas to be assessed. In general, the frameworks are updated prior to each new assessment to reflect the most current thinking in the field. The various frameworks for the 1990 and 1992 assessments are described below.*

NAEP's 1990 and 1992 **mathematics** assessment framework is a five by three matrix specifying five content areas--Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions--and three process or ability areas. These include conceptual understanding, procedural knowledge, and problem solving.

The objectives were developed under the auspices of the Council of Chief State School Officers (CCSSO) through a special NAEP Planning Project sponsored by the National Center for Education Statistics (NCES) and the National Science Foundation. This project involved widespread participation and review, including an objectives committee of mathematics educators; a steering committee with 18 members representing policy makers, practitioners, and citizens at large; distribution to the mathematics supervisors in the education agencies of all 50 states for review by state committees; and reviews by mathematics scholars, NCES staff, and NAEP's governing board--at that time, the Assessment Policy Committee (APC).

The 1992 **reading** assessment framework includes reading for three primary purposes--for literary experience, for information, and to perform a task. The process dimension includes fluency, constructing meaning (forming an initial understanding of the text and developing an interpretation of it), and elaborating and responding critically (reflecting on and responding to the text as well as demonstrating a critical stance).

Three types of materials form the basis of the assessment--literary and informational texts and documents.

NAGB began the consensus development process by soliciting written comments and holding public hearings to gather reviews of the 1990 reading objectives and suggestions for improving them. The updated framework was produced by the CCSSO under contract to NAGB. Achieving consensus on the objectives involved planning by a committee of reading experts guided by a steering committee with members recommended by 16 national organizations, as well as widespread review by practitioners in the reading field, assessment experts, school administrators, and state staff in both reading and assessment.

The 1992 **writing** objectives focus on students' ability to write effectively for a variety of purposes, including informative, persuasive, and narrative writing. They also emphasize students' abilities to manage the writing process and to meet standards of organization, elaboration, and convention.

NAGB initiated the consensus development process by widely distributing the objectives from the previous assessment (1988) and soliciting written comments about how they could be improved. The updating was conducted under the guidance of a committee of writing and assessment experts, as well as writing educators and practitioners, with the assistance of NAGB staff. The reformulated objectives were reviewed by a cross-section of state education personnel, representatives of business and industry, and writing and assessment experts.

The 1990 **science** framework encompasses a matrix of three content areas--life sciences, physical sciences, and earth and space sciences--and three broad process areas or thinking skills--conducting inquiries, solving problems, and knowing science. The foundation of the matrix is understanding the nature of science.

Objectives for the 1990 science assessment were developed under a grant from NCES to ETS, which managed the consensus process. An Assessment Development Panel including science education and assessment experts from universities and professional associations, as well as from state and local education agencies, guided the process with the assistance of ETS staff. An iterative series of reviews was conducted by the item development panel, science education and assessment experts, science curriculum coordinators from schools and state education agencies, scientists, school administrators, and NAEP's governing board (then the APC).

#7. **Question: What are the assessment questions like? Do they keep pace with curricular advances?**

Answer: *The assessment questions are continually updated to measure a broad range of content within each subject area. Each new version of assessment objectives reflects changes in curriculum and instruction, and thus innovations are required in the assessment instruments to keep pace with these changes. Even when the objectives stay the same, as for the 1992 mathematics assessment, many of the questions are replaced to reflect current advances in content and instruction.*

Because NAEP also updates its assessment methods, the types of questions vary across subject areas. Questions are tailored to the particular demands of the subject-area content. For example, the mathematics assessment provides students with scientific calculators and asks for open-ended responses to complex problems. In 1992, the NAEP reading assessments will break new ground by presenting longer, naturally occurring passages and increasing the open-ended questions to account for 40 percent of the total assessment. At grade 4, there will also be an oral reading assessment and a portfolio study. Based entirely on student writing samples, the writing assessment includes a variety of prompts addressing different writing purposes, and the students' responses are evaluated for task accomplishment, overall fluency, and mechanical correctness. The science assessment also includes a variety of open-ended questions, including written descriptions of students' conceptions of scientific inquiry and their drawings of various scientific phenomena and events.

The various 1990 and 1992 assessments, as detailed below, are a direct outgrowth of the consensus objectives and specifications developed for each subject area. Just as each subject area requires different item types, these in turn often require different testing designs.

Mathematics

At each grade level, both the 1990 and 1992 mathematics assessments include ten different 15-minute segments or "blocks" of multiple-choice and open-ended content questions. As part of the objectives development process for each subject area, the cells in the assessment frameworks (content by process matrices) are assigned weightings. Thus, the mathematics assessment questions were developed and assembled into the blocks--with each assigned both a content and a process classification--to reflect the weightings of the cells in the five by three matrix comprising the mathematics framework. Two of the ten blocks are designed to be answered using a calculator and three are presented accompanied by a paced audiotape to assess students' estimation skills and provide for complex problem-solving situations where students might become bogged down. Because the blocks contain a variety of item types, there are no rigid criteria dictating parallel structure across blocks.

The blocks are assembled three to a booklet and each student is asked to respond to one booklet. The three blocks accompanied by audiotape are assembled into one booklet, while the remaining seven blocks (including the two requiring calculators) are balanced across seven booklets (see #10).

The 1990 mathematics assessment contained 143 questions at grade 4, with 41 of them requiring open-ended responses. At grade 8, there were 191 questions and 42 of them were open-ended. Similarly, twelfth graders were administered a total of 203 questions, of which 47 were open-ended. The eighth-grade state assessments were based on seven of the 10 blocks, including the two requiring

scientific calculators, but not the three in the special study booklet requiring tape recorders. The seven blocks included 137 questions, of which 35 required open-ended responses.

NAEP's 1990 national mathematics assessments of fourth, eighth, and twelfth graders included more than 100 distinct open-ended questions (many administered at more than one grade), each with its own different scoring criteria. These items combined with the data collected in the state assessments yielded approximately two million open-ended responses.

For 1992, the actual numbers of mathematics questions are not yet known, but the total will be very similar to 1990, because the new assessment will also contain ten 15-minute blocks of questions at each of three grade levels. NAEP will use an innovative procedure to update the assessment, whereby four new replacement blocks (three unpaced and one paced by audiotape) will be developed at each grade level. It is anticipated that some of the questions from the 1990 special study audiotaped blocks measuring complex problem solving can be incorporated into unpaced blocks in 1992.

Several new blocks are field tested for each block that is to be replaced. The block that performs best in the field test is designated as the "parent block" and changed as little as possible, only with questions incorporated from the other field-tested materials. The characteristics of the new blocks will be parallel to those included in the 1990 assessment, thus, simplifying the subsequent scaling activities. This procedure provides for updating almost half the assessment in a forward-looking direction and NAEP currently field tests every other year in alternating years with the assessments. However, if the assessments were conducted every year, with fewer subjects in each year, then field testing could be conducted in conjunction with the assessments and a steady level and pace of work could be achieved.

The trend assessment directly linked to past mathematics assessments will be identical in both 1990 and 1992. It contains 127 multiple-choice and 34 open-ended questions at age 9, 158 multiple-choice and 27 open-ended questions at age 13, and 231 multiple-choice and 56 open-ended questions at age 17. Replicating procedures established in 1973 and used in each mathematics trend assessment since then, these materials are administered using a paced audiotape at all three age levels.

Reading

Plans for NAEP's new 1992 reading assessment include six 25-minute content blocks at each grade level, each containing one relatively long passage and approximately 15 open-ended and multiple-choice questions. At grade four, three of the blocks will focus on reading for literary experience and the other three on reading for information. At grades 8 and 12, two of the blocks will focus on each of the three purposes for reading--literary, informational, and task-oriented (documents). In addition, at grades 8 and 12 there will be two 50-minute blocks, one containing a long literary piece and the other requiring a comparison between two informational pieces. The long blocks may contain as many as 25 questions.

Although the number of questions will vary by block according to the reading purpose being measured, at least one of the open-ended questions in each block will require a paragraph-length response. The blocks will be assembled into booklets, either containing one long block or two of the six 25-minute blocks, with each student responding to one booklet.

At grade 4, a subset of the students will be asked to participate individually in an oral reading assessment, where for one of their assessment blocks, they will be asked to read the passage aloud and then respond to the questions in an interview mode. Their responses will be tape recorded and their oral reading

analyzed for fluency, by both timing the length of the response and counting the number of miscues. Miscue analysis involves coding the partial words, substitutions, omissions, insertions, reversals, and repetitions in students' oral responses across an inventory of features, including dialect, intonation, grammatical function, semantic function, and meaning change. In addition, NAEP will also pilot a reading portfolio study at grade 4, perhaps for the same students that participate in the oral reading study. Thus, for a national sample of students, NAEP will have an extensive inventory of students' ability to respond to text--oral reading and interview responses to questions for one passage, silent reading and written responses to six different passages, and examples of classroom work produced in response to novels, plays, and other works. This information will be further supplemented by teacher questionnaire data (see Question 9).

Although the fourth-grade trial state reading assessments will include the same six 25-minute blocks and teacher questionnaire information as the national assessment, the oral reading and portfolio assessments would involve an additional cost for participating states. Perhaps if these innovative assessment strategies are successful at the national level, they can be incorporated as regular features of future state assessments.

As is done for all NAEP assessments, the questions for the 1992 reading assessment will be developed--each with a content and process classification--to reflect the weightings designated in the assessment framework and specifications.

In comparison with the 1992 reading assessment, the 1990 reading assessment--which was based on the same underlying concept of reading, but is not as elaborate--was based on seven 15-minute blocks of passages and questions at each grade level. Fourth graders were asked to respond to 83 questions (two of them open-ended) pertaining to 24 brief passages, 11 literary and 13 informational. Eighth graders were asked to respond to 100 questions pertaining

to 27 brief passages (6 literary and 21 informational) and two-full page advertisements. Twelfth graders were asked to read the same advertisements and 26 passages (4 literary and 22 informational) and respond to 110 questions. At grades 8 and 12, only one of the questions was open-ended.

The 1992 reading trend assessment linking to the past will be identical to that conducted in 1988 and 1990. Based on literary and informational passages, most questions ask students to read for specific information or general understanding. This assessment includes ten 15-minute blocks of reading passages and questions at each of the three age levels. The assessment administered at age 9 includes 54 passages and 118 questions, including five open-ended. Thirteen-year-olds are asked 94 questions, six of them open-ended, pertaining to 40 passages. Seventeen-year-olds are asked 112 questions, nine of them open-ended, pertaining to 34 passages.

Writing

Plans for the 1992 writing assessment include six 25-minute blocks at each grade level, each containing only one prompt. In addition, two 50-minute prompts (one narrative and one informative) will be administered at grade 8 and three 50-minute prompts (one informative, one persuasive, and one narrative) will be administered at grade 12. In contrast, the 1988 writing assessment included seven 15-minute and three 30-minute prompts at each grade level.

The 1990 writing portfolio study involved collecting examples of student writing produced under classroom situations where students could have the advantage of peer and teacher review as well as the time to use external resources and engage in time consuming planning, drafting, revising, and editing activities. The teachers of fourth- and eighth-grade students were asked to provide NAEP with examples of students' best writing. Plans are to continue the portfolio assessment in 1992, because this assessment strategy permits NAEP to expand its description of students' test writing to include descriptions of their best writing.

The writing trend assessment consists of six 15-minute writing prompts at each grade level--4, 8, and 11. Results are produced for single prompts and on a scale across prompts. Detailed analyses of grammatical and mechanical errors are also conducted to measure trends in students' understanding of the conventions of written English. For the 1990 trend assessment, student performance also can be linked to their portfolio papers.

Science

The questions in the 1990 science assessment were written by science teachers, science educators, and test development staff. Each question in the assessment was classified according to a content and thinking-skill matrix that represented a consensus-derived science framework. The assessment consisted of seven 15-minute blocks of cognitive questions at each grade. The blocks were assembled three to a booklet and each student was asked to respond to one booklet. Each booklet consisted of a blend of questions that represented a specified balance across science content and skill areas.

For each grade, about 80 percent of the assessment time was devoted to multiple-choice questions and about 20 percent of the time was devoted to open-ended questions. About two-thirds of the open-ended questions were "figural response," whereas the other one-third were essay or short written answer. The figural response questions required students to mark or draw responses to indicate direction, location, or arrangement of objects and they required students to interpret and graph data. These types of questions are amenable to hand or machine scoring.

The 1990 science assessment contained 112 questions at grade 4, 146 questions at grade 8, and 150 questions at grade 12.

#8. **Question: What efforts are made to ensure that the NAEP achievement measures are not biased against any population groups?**

Answer: *Considerable energy and resources are expended to protect against item bias in the NAEP assessments. During the development process, each item is reviewed by trained ETS sensitivity reviewers in addition to measurement and subject matter specialists as well as experienced editors. External reviewers, including state education agency personnel, also review the items for appropriateness across regions and for students from a variety of backgrounds. In addition, NAGB is responsible for ensuring that all items selected for use in NAEP are free from racial, cultural, gender, or regional bias. NCES also reviews all NAEP items, and the background questions are subjected to further review by the Office of Management and Budget. As a final quality control precaution to monitor against bias, the actual item-level results are checked empirically.*

ETS is committed to ensuring that its tests acknowledge the multicultural and multiethnic nature of our society. More specifically, to support the many reviews, ETS, the principal operations contractor, has pioneered methods of protecting against item bias.

- Every item on every ETS-developed test--including NAEP tests--is subjected to mandatory "sensitivity review" by specially trained staff following established guidelines and procedures for rooting out biased, stereotyped, ethnocentric, elitist, or controversial material. These reviewers also check for balance across the pool of assessment questions.

- ETS applies a statistical procedure it developed in 1987--Differential Item Functioning (DIF) analysis--to all its tests. DIF analysis permits comparison of item performance across racial/ethnic and gender groups of similar demonstrated ability; it helps identify potential sources of bias for further investigation. DIF analysis are performed simultaneously with other statistical analyses, prior to scaling, and do not delay the reporting of results.

#9. **Question: What contextual background data does NAEP provide to help decision makers interpret the achievement results?**

Answer: *NAEP collects information from students, teachers, and school principals about hundreds of contextual background variables related to student, teacher, and school characteristics as well as curriculum and instruction. In developing the background questions, NAEP ensures that the questions do not infringe on respondents' privacy, that they are grounded in research, and that the answers can help inform the debate about educational reform. In 1990 and 1992, NAEP plans to relate all background information, for both the nation and participating states, to the student performance data.*

The following describes explains the process for developing the 1990 and 1992 background questionnaires and summarizes their contents.

In coordinating NAEP background data collection with other projects--such as the Schools and Staffing Survey, the National Education Longitudinal Study, and the National Cooperative Education Statistics System/National Forum on Education Statistics--NCES obtains guidance from NAGB and its committees and gathers information from a variety of sources within and external to the government. This information is supplied to a committee of experts on educational indicators, along with data on the validity and utility of particular questions used in previous assessments and a contractor review of current research literature in teaching and learning. The committee then drafts the background-question framework for each assessment, which is circulated for widespread review and revised accordingly. For both the 1990 and 1992 assessments, the committee work was staffed by ETS. For the 1990 assessment, the resulting framework focused on six educational areas: curriculum, instructional practices, teacher qualifications, educational standards and reform, school conditions, and conditions outside of the school that facilitate learning and instruction. For 1992, the framework currently reflects four broad areas: instructional content, instructional practices and experiences, teacher characteristics, and school context and conditions.

To reflect the most current thinking in the field, the framework for NAEP background questions is revised and updated with each assessment. Some core questions are retained across assessments to measure trends, while some outdated or less useful questions are replaced to keep NAEP abreast of new research findings and educational reforms.

In 1990, NAEP administered the following questionnaires to students, teachers, and principals. For efficiency in data collection and to link student background and achievement data, the two student questionnaires were included in the assessment booklets and administered in the same session as the tests. With the exception of fourth graders, students are given five minutes to respond to the questionnaires. To improve the validity of fourth graders' answers to the questionnaire about their demographic backgrounds, the assessment administrators read the questions aloud and clarified the vocabulary and intent of the questions.

- The **student demographics questionnaire** (18 questions at grade 4, 21 questions at grade 8, and 30 questions at grade 12) includes questions about race/ethnicity, language spoken in the home, mother's and father's level of education, reading materials in the home, homework, attendance, school climate, academic expectations, which parents live at home, and which parents work. This questionnaire is included in every booklet.

- The **student questionnaire about educational experience in the subject area** (ranges from 14 to 35 questions, depending on the grade level and subject area) includes questions about instructional activities, courses taken, use of specialized resources such as calculators in mathematics class, and views about the utility and value of the subject matter. This questionnaire is specific to each subject area. For example, the mathematics booklets have questionnaires on mathematics instruction and course-taking, and the reading booklets have questionnaires about reading instruction and experiences.

To supplement the information on instruction reported by students, the teachers of students participating in NAEP are asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. This information is linked directly to their students' performance data. In 1990, questionnaires were administered to the teachers of fourth and eighth graders participating in the mathematics assessments, **including the teachers of all the eighth graders who participated in the trial state assessment.** In addition, administration of an eighth-grade science teacher questionnaire was made possible by supplemental funding from the National Science Foundation through a subcontract from Horizon Research, Inc. to the NAEP operations contractor (ETS).

In 1992, plans are to administer questionnaires to all the teachers of students participating in the fourth-grade reading and mathematics assessments and the eighth-grade mathematics assessments, **including the teachers of all students participating in the trial state assessment.** Supplemental funding may be sought for an eighth-grade writing teacher questionnaire that could track recent events in the instruction of process writing and also provide trend information from the writing teacher questionnaire administered as part of the 1988 writing assessment.

The teacher questionnaires contain two parts. The first part pertains to the teachers' background and training. The second part pertains to the procedures she or he uses class by class for each class containing an assessed student.

- The **Teacher Questionnaire, Part I: Background and Training** (at grades 4 and 8, 34 questions for mathematics; at grade 8, 100 questions for science) includes questions pertaining to gender, race, ethnicity, years of teaching experience, certification, degrees, major and minor, coursework in education, coursework in subject area, in-service training, extent of control over classroom, instruction, and curriculum, and availability of resources for classroom.

- The **Teacher Questionnaire, Part II: Classroom by Classroom Information** (at grade 4, 34 questions; at grade 8, 35 questions for mathematics and 58 questions for science) includes questions on the ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, opportunity for students to learn (instructional emphasis given to) the topics and skills covered in the assessment, and use of particular resources.

The School Characteristics and Policy Questionnaire is given to the principal of each school that participates in NAEP. Including the schools in the trial state assessment, more than 5,000 school questionnaires were administered in 1990. Plans are to continue this practice in 1992, which will include administering nearly 12,000 questionnaires.

- The **School Characteristics and Policies Questionnaire** (at grades 4 and 8, contains 117 questions and at grade 12, contains 125 questions) collects information about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, size and composition of teaching staff, policies about tracking, curriculum, testing practices and use, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

It is NAEP's intent to assess all selected students. Therefore, all selected students who are capable of participating in the assessment should be assessed. However, some students sampled for participation in NAEP are excluded from the sample according to carefully defined criteria. Specifically, some of the students identified as having Limited English Proficiency (LEP) or having an Individualized Education Plan (IEP) may be incapable of participating meaningfully in the assessment. These students are identified as follows:

LEP students may be excluded if:

- The student is a native speaker of language other than English; AND
- He or she has been enrolled in an English-speaking school for less than two years; AND
- The student is judged to be incapable of taking part in the assessment.

IEP students may be excluded if:

- The student is mainstreamed less than 50 percent of the time in academic subjects and is judged to be incapable of taking part in the assessment, OR
- The IEP team has determined that the student is incapable of taking part meaningfully in the assessment.

When there is doubt, the student is included in the assessment.

For each student excluded from the assessment, including those in the trial state assessment program, schools are required to complete a questionnaire about the characteristics of that student and the reason for exclusion.

- The **Excluded Student Questionnaire** collects data about students' race/ethnicity and the reason for exclusion from the assessment. For IEP students, the questionnaire includes questions about students' functional grade level, mainstreaming, and special education programs. For LEP students, it asks about students' native language, time spent in special education and language programs, and the level of the students' English language proficiency.

Typically, about five percent of the students are excluded from the national assessments. The information from the Excluded Student Questionnaires will be particularly useful in providing information about differential exclusion rates across states participating in the trial state assessment.

#10. **Question: Does NAEP use matrix sampling to reduce the burden for participating students? What is "Focused-BIB Spiraling" and what are the advantages of using it in NAEP?**

Answer: *To reduce the burden for students, NAEP uses a powerful variant of matrix sampling--Focused-Balanced Incomplete Block (BIB) Spiraling. In matrix sampling, the total pool of assessment questions is divided, and portions are given to different but equivalent samples of students. Thus, not all students are asked to answer all questions. This system provides broad coverage of the subject being assessed while minimizing the classroom time required of any one student. NAEP samples enough students--about 30,000 per subject area--to obtain precise results for each question, while each student invests only about an hour in the assessment--10 minutes on background questions and 45 or 50 minutes on test questions. Only a small proportion of students participate in NAEP (at the most, only about two-hundred thousand of the three and one-half million students at any grade level).*

More specifically:

- The "focused" part of NAEP's matrix-sampling method means that each student responds to questions from only one subject area being assessed.

- The BIB part of the method ensures that some students receive interlocking parts of assessment forms, enabling NAEP to check for any aberrant interactions between the different samples of students and the different sets of assessment questions.

- Spiraling means that different assessment forms are distributed across the students sampled. Thus, students in any one assessment session receive a variety of forms, which reduces the likelihood of copying answers from their neighbors. Spiraling also enhances the equivalence of the samples of students responding to each form, thereby increasing the effectiveness of the matrix sampling technique.

This sampling method means that assessment questions are not "booklet-bound." For example, if a secondary analyst was interested in comparing answers across questions for individual students, all the pairs of questions are available for analysis. With simple matrix sampling, separate sets of questions are confined to particular booklets--a major roadblock to investigating items that cut across booklets. The 1990 and 1992 assessments also include some simple matrix sampling, because it suits the needs of certain assessment questions; however, the more sophisticated method is used for the majority of questions to produce more useful data.

Focused-BIB spiraling provides for discrete blocks of exercises to be arranged so that each pair of blocks occurs together in at least one booklet. Under this plan, a greater variety of booklets is printed than would be the case in an assessment based on simple matrix sampling, but the total number of questions is the same, as is the total number of students responding to each question. Thus, BIB spiraling does not add time to the assessment process once the booklets are printed and the scanning process is computerized to account for the different combinations of forms--both activities that are accomplished well before the data arrive from the field.

Two versions of Focused-BIB spiraling occur in the 1990 and 1992 assessments. The design used for the 1990 mathematics and science assessments follows a seven-block, seven-booklet design. As illustrated below, every pair of blocks appears in one booklet and each block appears in three booklets. In the NAEP assessments using this design, each block is 15 minutes long.

Booklet	Blocks
1	1 2 4
2	2 3 5
3	3 4 6
4	4 5 7
5	5 6 1
6	6 7 2
7	7 1 3

This design also provides for balancing the position of items across booklets. In a simple matrix design, the same items are always last in the booklet and subject to underestimating students' ability because of a fatigue factor. For the NAEP mathematics and science assessments, each block of items occurs once in each block position--first, second, or third. This means one-third of the students respond to these items first, one-third respond to these items in the middle of the booklet, and one-third respond to the items in the last part of the booklet. As a result of this balancing, NAEP has found that in assessments that rely more heavily on open-ended responses, including reading and writing, students generally do less well on the last block of questions, regardless of their content.

For the 1992 reading and writing assessments, each of the six blocks appears with every other block in one booklet and each block appears in six booklets, three times in the first position and three times in the second position. As illustrated below, this design requires 18 booklets.

Booklet	Blocks	Booklet	Blocks	Booklet	Blocks
1	1 2	7	4 2	13	6 5
2	2 1	8	3 4	14	1 6
3	3 1	9	5 1	15	6 2
4	2 3	10	2 5	16	3 6
5	4 3	11	5 3	17	6 4
6	1 4	12	4 5	18	5 6

It should be stressed that BIB designs may be readily developed for a wide array of sizes of item pools and types of questions.

#11. **Question: How are students selected for participation in NAEP?**

Answer: *Students are selected according to scientific procedures designed to yield nationally representative results as well as results for particular subpopulations of students, as defined by gender, race/ethnicity, region of the country, and size/type of community. For the trend assessments, NAEP assesses 9-, 13-, and 17-year-old students. More recently, NAEP also began assessing public and private school students at grades 4, 8, and 12, corresponding to the conclusions of primary, middle school, and high school experiences. For the 1990 and 1992 trial state assessments, only students attending public schools and in particular grades are assessed--eighth grade in 1990 and fourth as well as eighth grade in 1992.*

The details of the sampling procedures for the national and state assessments differ in some respects. For the national assessments, NAEP uses a deeply stratified, four-stage sampling design. The first stage involves stratifying primary sampling units (typically aggregates of contiguous counties, but sometimes a single county) by region and community type and making a random selection. Second, within each primary sampling unit, public and private schools are enumerated, stratified, and randomly selected. For each selected school, students are randomly selected. Finally, sampled students are assigned to the different kinds of assessment sessions, including paced-tape sessions for either the trend assessments or mathematics estimation, mathematics sessions requiring instructions on how to use calculators, reading and writing sessions with 25- and 50-minute blocks, and so forth. To increase sampling efficiency, a number of assessment sessions are conducted within each participating school. Thus, multiple subject areas are typically assessed within any given school. (For school and student sample sizes, see #5.)

For the states participating in the trial state assessments, the schools in each state are enumerated, stratified, and randomly selected. Within each school, students are listed and randomly selected. In 1990, students did not need to be assigned to different session types because all of them were participating in the newly developed mathematics

assessment. However, in 1992, fourth graders will be assigned either to the reading assessment, which has two 25-minute sections, or the mathematics assessment, which uses calculators and is comprised of three 15-minute sections.

For each state in the 1990 trial assessments, approximately 2,500 eighth graders were sampled from approximately 100 schools. The same procedures will be used in 1992, including some refinements of the general strategy used in small states that do not have 100 eighth-grade schools. For the fourth-grade reading and mathematics assessments, 2,500 students will be selected for each subject area in each state. In the interest of sampling efficiency, both reading and mathematics sessions will be conducted in selected schools with sufficient numbers of eligible students. However, because they tend to be small, some elementary schools will only have enough students for one session. Thus, approximately 125 fourth-grade schools will be sampled in each state.

It is NAEP's intent to assess all students selected for participation in national or state assessments. However, some students sampled can be excluded according to carefully defined criteria (see #9).

Having so many different NAEP samples--including the different subject areas, the trend assessments, and the different states--complicates the process of "weighting" the actual assessment results to account for differential probabilities of selection into the sample and to reflect the Census Bureau's population estimates for the nation and the various states. However, considering the burden and resources associated with the alternative--assessing every student in every subject--the eight weeks this process takes is time well spent. Further, this process does not affect the overall schedule for producing assessment reports, because initial data analysis and scaling activities (item calibration) occur simultaneously based on preliminary weights supplied well before the data arrive from the field.

#12. Question: Who ensures the cooperation of sampled schools and administers the NAEP assessments and questionnaires?

Answer: *All NAEP data are collected by trained administrators. For the 1990 and 1992 assessments, collecting the data for the national assessments--trend and newly developed--is the direct responsibility of ETS and its subcontractor, Westat, Inc. However, in accordance with the NAEP legislation, data collection for the trial state assessments is the responsibility of each participating state. Uniformity of procedures across states is achieved through training and quality control monitoring by Westat, Inc.*

For the national assessments, the contractor, Westat, trains its own field staff to collect the data, thus, reducing the burden on participating schools. However, because NAEP relies on the goodwill of school administrators, obtaining cooperation requires substantial time and energy. The sampled schools are notified of their selection through a series of mailings, including letters to the Chief State School Officers and district superintendents. In addition, informational materials are sent and procedures are explained at in-person introductory meetings.

Westat personnel also train local field staff for the state assessments. The NAEP legislation requires that participating states provide for the data-collection activities, including ensuring the participation of sampled schools and students, assigning quality personnel to assess students according to standardized procedures, distributing and collecting the questionnaires, and observing procedures for test security. In 1990, in addition to training the local administrators, Westat provided quality control across states by monitoring half the sessions in each state. Since the state data collection efforts seem to have been successful, it is anticipated that fewer sessions will be monitored in 1992--perhaps 20 percent of the sessions, rather than 50 percent.

In September, a five-day training session was held by Westat to orient and train the 40 National Supervisors assigned to the 1990 National Assessment of Educational Progress and the 40 Westat State Supervisors assigned to the trial state assessment.

Many of these individuals had supervised NAEP assessments in previous years and participated in the field test of the trial state assessment in the spring of 1989.

Additional training of 40 Westat State Supervisors and 160 Quality Control Monitors (QCMs or "NAEP Representatives") was done by Westat during the first week of January. All staff were trained in administrative and monitoring procedures. The State Supervisors also were trained to be trainers of the Local Administrators from each state. During January and into February, about 360 one-day training sessions for Local Administrators were conducted by the State Supervisors. Each training session took about six hours to cover all of the material. In all, about 4,000 Local Administrators were trained within a three-week period. To assure uniformity in the training sessions, Westat developed a 50-minute video training presentation, which was supplied to each supervisor and accompanied by a scripted trainer's guide and practice exercises. A Manual for Local Administrators also was prepared for use during the training and for the Local Administrators to use extensively during the actual assessment sessions.

The QCMs were trained to be able to act in the role of a Local Administrator as well as being able to monitor and evaluate the Local Administrator's tasks. One QCM was assigned to monitor one of the quadrants within a state (four monitors per state). When visiting an assessment within a school, the QCM was able to assist the Local Administrator and help in some of the responsibilities, e.g., post-assessment quality control, packing assessment materials, and arranging for the return of materials to NCS.

#13. Question: How does NAEP reliably score millions of open-ended responses without delaying the reports? How is the open-ended scoring merged with the computerized scoring of multiple-choice questions?

Answer: *Responses to the open-ended questions in NAEP assessments are evaluated by professional readers who are trained to use the scoring guidelines developed for each item. NAEP assembles a high-quality staff with expertise in the subject areas being scored and trains them as readers until they are in agreement on applying the scoring criteria*

uniformly. During the scoring, reliability reports are used to monitor accuracy, and corrective steps are taken as necessary. The open-ended scoring is conducted in advance of scanning, thus, the data for multiple-choice and open-ended questions are recorded in one pass.

To ensure that the scoring is conducted reliably and efficiently, NAEP trains the readers and continuously monitors their work using a set of well established procedures. The scoring effort includes the following steps.

- **Preparing for training**: Before scoring begins, subject-area specialists, scoring directors, and table leaders meet and read hundreds of student responses to each open-ended item. This activity allows those leading the scoring to become familiar with the range of performance and to refine the scoring guidelines as appropriate. During this meeting, training materials are prepared. First, sample responses are selected to illustrate each score point defined in the guidelines established for each open-ended item. Second, practice papers are selected for use in the training sessions.

- **Training the readers**: The training of readers is directed by subject-area specialists who review the scoring guidelines and sample responses, and then lead the readers in practice scoring.

- **Quality control monitoring**: Monitoring accuracy is a critical aspect of any successful scoring effort. During the initial scoring period, subject-area specialists and table leaders review student responses that have been scored. This allows them to check whether or not the readers are correctly implementing the scoring guidelines and to determine each reader's propensity for error-free scoring. Throughout the scoring, table leaders routinely read scored responses to monitor each reader's work. In addition, reliability reports are generated, providing precise quantitative information on the scoring, both by reader and by item. The scoring

director and table leaders use both the observational and quantitative information to identify and rectify any problems that may occur.

Measuring trends in writing achievement involves linking the results from the current assessment to the results from the previous trend assessment. It is possible that the ratings provided by the group of scorers assembled to score the current assessment will differ from the ratings of the scorers who were assembled to score the previous assessment. If uncontrolled, this between-year scorer effect adds a confounding factor that is detrimental to the measurement of trend in, for example, writing or reading ability.

To minimize differences in scoring between two successive assessments, scorers are trained using a selection of papers given in the previous assessment. However, experience indicates that, despite consistent score standards and extensive training, there will be some difference in the scoring patterns between successive assessments. Thus, a random sample of 20 to 25 percent of the responses from the prior assessment is systematically interspersed among the current responses for rescoring. The results are then used to determine the degree of scoring agreement between the current and previous assessments. If necessary, current assessment results are then adjusted to account for any differences. The differences can lie in the relative stringency of scoring, or in the variance of the scores, or both. Additionally, the degree of agreement in scoring patterns between successive assessments can differ by item. For example, the procedure developed by NAEP to adjust for differences in scoring patterns between successive writing assessments is sufficiently flexible to allow for item-specific differences in the mean and/or standard deviation of the scores.

The time required to conduct open-ended scoring is primarily a function of the amount of material to be scored and the number of people used to accomplish the task. Obviously, more people will get the job done more quickly, although some more time must be spent on training large numbers of people to score uniformly, especially since NAEP scorers must evaluate responses according to defined categories and not

"holistically." For the 1990 assessment, which generated 3.1 million open-ended responses, more than 77 readers were used, including scoring directors and table leaders, to complete the open-ended scoring one month after final data collection.

NAEP incorporates several time-saving devices and efficiencies into its open-ended and machine scoring operations. For example, materials collected in the last week of the field administration are returned to the scoring contractor (NCS) by overnight mail. In addition, open-ended responses are scored prior to scanning so that all materials can be scanned in one pass and the data for open-ended and multiple-choice items do not have to be merged later.

In 1990, NAEP initiated the use of bar codes on the covers of assessment booklets, which increased the accuracy and speed of processing by allowing machine-scoring of student information that was previously gridded by hand. It is estimated that this state-of-the-art technology increased the accuracy of processing to 99.9 percent or better. Further, by eliminating the need for assessment administrators to transcribe student identification numbers on the booklet covers, student confidentiality is better protected. NAEP will continue to use the bar code technology in the 1992 assessment.

For 1992, in order to provide the trial state assessments reports in timely fashion, all the scoring for the 1992 reading and mathematics materials will be accomplished three weeks after the data are collected. Although many more materials will be collected than in 1990 for the eighth-grade mathematics assessment, the required systems will not have to be developed from scratch as in 1990, and they can be modified to accommodate the new materials prior to the receipt of the 1992 data.

#14. Question: How does NAEP analyze the assessment results?

Answer: *The analysis of NAEP data are accomplished in two major phases. The first phase, based on Item Response Theory (IRT) scaling methods, involves obtaining the achievement results for the various subject areas assessed, and for the trial state program,*

equating national results and the state results. The second involves analyses based on the achievement scores, such as relating the achievement results to the numerous background variables.

The procedures used by NAEP to conduct the thousands of analyses required, as determined by the reporting requirements of the various national and state assessments and by the large number of background variables associated with each assessment, have been developed because they are sensitive to limiting student burden, take the least time to complete, and provide the most accurate results. Finally, these procedures provide data readily available for use by secondary analysts.

Because of the importance of the data, both in terms of the amount of money expended to obtain it as well as the ever increasing visibility of the reports based on this data, the scaling and analysis of NAEP data in general, and of the trial state assessment data in particular, are conducted in a careful manner that includes extensive quality-control checks.

The following steps are used to generate scale-score data files suitable for analysis:

- Immediately after receipt of the machine-readable data tapes containing students' responses, all cognitive and noncognitive items are subjected to an extensive item analysis to assure that each item represents what it is purported to measure. The results are reviewed by knowledgeable project staff in search of anomalies that might signal unusual results or errors in creating the database. In parallel with this item analysis, each cognitive item is examined for indications of Differential Item Functioning (DIF). All steps in these analyses for the 1990 and 1992 assessments will use preliminary weights when appropriate to approximately account for the disproportionate representation of certain subgroups in the sample. Thus, the analyses are not delayed for lack of the final weights. By using the

preliminary weights, these essential quality-control steps can be conducted as early as possible.

- After completion of the item analysis, preliminary estimates of the parameters of the IRT model are obtained for each item. This item calibration also uses the preliminary weights and results in a preliminary subscale for each of the predefined content areas specified in the assessment frameworks. (For mathematics in 1990 and 1992, there are to be five subscales, each corresponding to a particular content area. For reading in 1992, there are to be three subscales.) To verify the subscale definitions, post-hoc dimensionality analyses are conducted, in parallel with the item calibration. Because the item parameters determine the representation of each item in the subscales, careful checks are made by the psychometric staff to verify that the IRT scaling model provides an acceptable representation of the student responses to the items. In particular, the fit of the model is examined, by item, for major demographic subgroups and, for the trial state assessment, for each state. Because data collection for the trial state assessments differed from that for the national assessment, item-parameter estimation is performed separately for state data and national data.

- Item calibration is conducted in parallel with the development of the final student sampling weights. Because of their importance to the final estimates, the weights are carefully checked to ensure that the weighted population totals are consistent with other available population information and that no student or group of students has undue impact on parameter estimates. After calculation of the final student weights, the final IRT item parameters are updated for each subscale, and final checks on the fit of the IRT model are made.

- The next step in the scaling process is to calculate plausible values of subscale proficiency scores for each student participating in the assessment. The plausible-values technology, which also uses student background information, allows for **more accurate estimates** of the performance of subpopulations and more appropriate estimates of the variability of those estimates than does the standard (and much simpler) procedure of estimating a standard proficiency score for each student based only on responses to the items. Development of the plausible values begins immediately after the preliminary item calibration and occurs parallel to the final item calibration, using the preliminary weights. After determination of the final item parameters, the final plausible values are generated using the final parameters and student weights. A series of careful quality-control steps are taken in constructing the plausible values to ensure the accuracy of subpopulation estimates based on these plausible values. The construction of plausible values is conducted separately for each state participating in the trial state assessment and for each national sample.

- The final step in scaling the trial state assessment results is linking the results for the state assessments to those for the nation. The selection of the linking function (which is anticipated to be linear) and its adequacy will be determined by comparisons of the distribution of scores for the two types of assessments.

With all the steps involved in scaling, including the careful quality-control checks, the entire mathematics scaling process for the 40 or more state assessments and for the nation is slated to take three-and-a-half months in 1992. The schedule is longer for 1990, because this is the first scaling ever attempted by NAEP that involves 40 different states as well as the nation, and because of the need to wait until the end of May for the national data. The schedule for scaling the 1992 reading data will also be longer, because that scaling will be the first attempted on the new reading items, which differ in

many important respects from those in previous assessments. The faster schedule for the 1992 mathematics scaling is possible because of its high degree of similarity to that performed for the 1990 assessment.

For the full array of NAEP data, the plausible-values approach takes less time than would numerous single runs of a specialized program estimating the proficiency scores of a single subgroup. In the specialized procedure, distributions of performance for a subpopulation can be obtained in narrowly defined content areas, without estimating scores for individual students. However, although NAEP could estimate proficiency distributions without generating plausible values, this would have to be done separately for each subpopulation in each report. Plausible values extend the specialized methodology to handle not just one subpopulation at a time, but all the potential interrelationships among proficiency scales and background variables. The plausible-values approach solves the estimation problem once--albeit with more work than any one or two or even ten of the simpler single runs--and permits completion of the hundreds of analyses required by the extensive number of NAEP background variables in less time than conducting practically endless separate estimations.

For the NAEP data, plausible values also provide more accurate estimates of student performance. In conjunction with subscales, plausible values allow for accurate and statistically unbiased estimates of population characteristics despite changes in test lengths, difficulties, and balances of item content. They are born of a statistically rigorous approach to handling multiple sources of uncertainty in data, and they are sufficiently flexible to accommodate the evolution of assessment envisaged for NAEP.

When the procedures and test items are essentially constant across administrations, as in traditional standardized testing programs, the error structure underlying the test also remains constant and simpler estimation procedures can be used. Because error in the test will be the same in each administration--for example, error in measuring differences in performance between girls and boys--any increases or decreases in gender differences can be assumed to be real and not a function of the test.

NAEP, however, continues to provide relevant information by changing its tests and needs to use methodology that can accommodate substantial updating from assessment to assessment and be sensitive enough to measure actual changes in student performance. Plausible values are currently the best way to control statistically, after the fact, for the variations caused by these differences, which are controlled operationally in other more standard programs.

It should be additionally noted that plausible values increase accuracy even for group-level reporting on an overall scale. In particular, even when students are administered sufficient items to use simpler procedures, some attempt very few items. Furthermore, this differential nonresponse is generally related to ability, so that relatively more lower ability students attempt fewer items. Plausible values help account for this situation and yield more accurate estimates of student achievement that, in turn, lead to more accurate estimates of the achievement of various subpopulations.

Finally, as a natural by-product, the same student-level data provided by the plausible-values approach, and upon which the NAEP reports are based, can be used by secondary researchers, who can also carry out the full range of NAEP analyses. The specialized approach, in contrast, does not yield detailed student-level information for use by a broad range of secondary researchers relying on standard statistical packages.

In summary, the NAEP implementation of IRT analysis provides, in an efficient way, for extensive, detailed analyses by NAEP staff and by secondary analysts of data that are not biased and have been subjected to numerous quality-control steps.

The plausible-values scaling technology is at the heart of NAEP's ability to perform the second phase analyses and report the type of results contained in, for example, The Mathematics Report Card: Trends and Achievement Based on the 1986 National Assessment (1988) and Learning to Write in Our Nation's Schools: Instruction and Achievement in 1988 at Grades 4, 8, and 12 (in publication). That is, it permits analyzing and reporting trend and baseline results for:

- Subpopulations of students defined by race/ethnicity, region of the country, size and type of community, gender, level of parents' education, type of curriculum, mode of instruction as reported by their teachers, and so forth (see response to #9).

- Student performance on the specific content areas within each subject area as defined by the consensus assessment frameworks (e.g., reading for literary experience, to be informed, or to perform a task; or, for mathematics, numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions).

- Comparisons of how students' performance improves as they progress through school, sometimes differentially for particular content areas or subpopulations. For example, in some subjects including science and U.S. history, gender differences favoring males increase as students move through school. In mathematics, students appear to learn arithmetic knowledge and procedures at a rapid pace in elementary schools, but growth in their skills and knowledge of the more complex mathematical content (i.e., geometry and algebra) taught in the middle grades and high school is much slower.

In addition, for the trial state assessment, analyses will be conducted to aid in the evaluation of the results, to compare distributions of achievement across states, and to relate performance with student, teacher, and school variables. The differences between monitored and nonmonitored sessions will be evaluated to verify that the results from the two administrative conditions can be legitimately combined. Subsequent to scaling, the complete analysis of the trial state mathematics assessment data is slated to take two months in 1990 and six weeks in 1992.

15. **Question: How can NAEP report results in a timely manner when procedures, and thus the computerized systems designed to implement those procedures, change from assessment to assessment?**

Answer: *NAEP uses computerized systems as much as possible, for example, in initial quality control steps to check data consistency as well as throughout all phases of the scaling and data analysis. A computerized report generation system has been developed to produce state reports similar to those prepared by hand for the national results. However, responding to the consensus process requires changes from assessment to assessment and these changes require new and revised systems. Thus, to prepare the new systems prior to receipt of data, ETS has developed a data simulator which is used to modify the systems and create new ones while the booklets are being printed and the data collected. By using the simulator in conjunction with procedures developed in 1990, NAEP anticipates having the reports of the 1992 national and state mathematics assessments available in 1992.*

The continued introduction of innovations into NAEP--new objectives, items, data collection procedures, analysis methods, and reporting technologies--precludes using static computer systems that can run unsupervised, mechanically-oriented data analyses. Even in 1992, when NAEP expects the mathematics assessment to closely parallel the 1990 assessment and a production system approach can be used to accelerate a large portion of the data analyses, changes will still occur. Thus, to speed up the reporting of NAEP results, a data simulator has been developed by ETS and used experimentally in the 1990 data analyses and is expected to be used extensively with the 1992 NAEP data.

When each assessment booklet, questionnaire, or other data form is finalized--before the data are collected from students, teachers, and principals--a set of simulated data is produced in identically the same format as the data that will be produced when the actual assessment takes place. The properties of the simulated data are decided by NAEP statisticians and thus are known, although not by the analysts and system operators. For example, simulated data can be generated to show greater or lesser group differences or rises or falls in performance trends. The simulated data are then

processed through all assessment procedures as a check on the computer programs and analytic techniques. Since the properties of the simulated data are in actuality known in advance, the results of the "dry run" analyses can be verified against the solutions that should have been obtained. Based upon this comparison, NAEP can detect any failures in the computer programs or statistical routines and take corrective action before the data arrive from the field.

The simulator process is being developed and tested before the NAEP 1990 data are processed. It is expected to result in some time-savings for the 1990 reporting schedule and in substantial savings (about two months) in 1992, when the properties of the state mathematics assessment data are better known and the simulator itself is perfected. These savings will enable NAEP to have camera-ready copy of the state mathematics assessment results on December 15, 1992--in the same year as the data collection.

However, even with the simulator, NAEP's accuracy and integrity require competent reviews of the results obtained from each step in scaling and data analysis. Experience has shown that with NAEP, the wise and prudent course is to expect the unexpected.

#16. Question: What are the NAEP scales?

Answer: *The NAEP scales are designed to provide a basis for describing student performance in different subject areas across grades, subpopulations, and assessment years in a way that can be easily understood. They place student performance on a common 0 to 500 metric that can be used to trace growth in achievement through the school years and across time. To give meaning to the results, student performance is characterized at various levels along the scale. Since 1984, when NAEP started using this technique for reporting student achievement policy makers at all levels have become increasingly aware of the value and utility of NAEP data.*

NAEP has used similar procedures in developing IRT scales for reading, mathematics, science, U.S. history, and civics. To "anchor" the scales, NAEP provides descriptions of student performance at various levels (a process that takes only a few days). Scale anchoring begins with empirical procedures whereby NAEP delineates sets of items that discriminate between adjacent levels of performance--that is, items likely to be answered correctly by students performing at one level on the scale and much less likely to be answered correctly by students performing at the next lower level. Then, the sets of items represented at each level are analyzed by panels of experts, who carefully consider and articulate the types of knowledge, skills, and reasoning abilities demonstrated by correct responses to the items in each set. This information is subsequently placed in the context of the assessment frameworks and used to characterize student understanding.

The scale descriptions for each of the five curriculum areas follow. Although proficiency levels, in theory, can be defined anywhere along the 0 to 500 scales, so few students perform at the extreme ends of the scales that it is not practical to do so. For reading, mathematics, and science, NAEP defined performance at five levels--150, 200, 250, 300, and 350. Level 150 was not described on the U.S. history and civics scales, because student performance did not vary as much in those subject areas.

By reporting the percentages of students who attain each level of performance across grades and over time, NAEP makes it possible to chart patterns and changes in American education. For example, it appears that the significant gains in reading achievement from 1971 to 1988 occurred at the lower end of the scale. Since younger students made gains during the 1970s and older students during the 1980s, it also may be that the recent improvements in reading achievement at age 17 are due, at least in part, to an early advantage.

The reading and mathematics scales will be used to report the trend data from the 1990 and 1992 trend assessments, and the science scale will be used to report the science trend data as well as the results of the 1990 assessment.

However, new scales will be developed for reporting the 1990 and 1992 mathematics assessments and the 1992 reading assessment, including the results obtained from the trial state assessments. The 1990 mathematics scaling may encompass several approaches, including a new version of the grade-comparative scale anchored in accordance with the new assessment items reflecting higher-order problem-solving skills and calculator usage, as well as an effort by NAGB to set goals for achievement. Because students at all three grade levels (4, 8, and 12) are assessed during the same time period and the data arrive from the field simultaneously, scaling the three grades together or separately takes about the same amount of time and provides equally accurate results.

The approach of combining information across grades provides for making policy relevant comparisons and for a more concise reporting strategy that is especially important considering the expansion of the trial state assessment in 1992--an additional grade, an additional subject, and the availability of the first trend data. Based on procedures used for the NAEP mathematics scales in 1990, decisions will be made on the approaches to be used for 1992, particularly for the new reading assessment.

For writing, NAEP has used a different procedure, called the Average Response Method (ARM), to develop scales. In ARM, students' papers are evaluated according to criteria that define unsatisfactory, minimal, adequate, and elaborated responses. The results are then scaled across tasks and grades. The writing trend results collected in 1990 and 1992 will continue to be reported on this scale. For the new 1992 writing assessment, however, NAEP will use a categorical item-response model. This will permit using an IRT-type scale that accommodates the full range of responses to each writing prompt in more flexible ways than in the past.

LEVELS OF MATHEMATICS PROFICIENCY

Level 150--Simple Arithmetics Facts

Learners at this level know some basic addition and subtraction facts, and most can add two-digit numbers with regrouping. They recognize simple situations in which addition and subtraction apply. They also are developing rudimentary classification skills.

Level 200--Beginning Skills and Understanding

Learners at this level have considerable understanding of two-digit numbers. They can add two-digit numbers, but are still developing an ability to regroup in subtraction. They know some basic multiplication and division facts, recognize relations among coins, can read information from charts and graphs, and use simple measurement instruments. They are developing some reasoning skills.

Level 250--Basic Operations and Beginning Problem Solving

Learners at this level have an initial understanding of the four basic operations. They are able to apply whole number addition and subtraction skills to one-step word problems and money situations. In multiplication, they can find the product of a two-digit and a one-digit number. They can also compare information from graphs and charts, and are developing an ability to analyze simple logical relations.

Level 300--Moderately Complex Procedures and Reasoning

Learners at this level are developing an understanding of number systems. They can compute with decimals, simple fractions, and commonly encountered percents. They can identify geometric figures, measure lengths and angles, and calculate areas of rectangles. These students are also able to interpret simple inequalities, evaluate formulas, and solve simple linear equations. They can find averages, make decisions on information drawn from graphs, and use logical reasoning to solve problems. They are developing the skills to operate with signed numbers, exponents, and square roots.

Level 350--Multi-step Problem Solving and Algebra

Learners at this level can apply a range of reasoning skills to solve multi-step problems. They can solve routine problems involving fractions and percents, recognize properties of basic geometric figures, and work with exponents and square roots. They can solve a variety of two-step problems using variables, identify equivalent algebraic expressions, and solve linear equations and inequalities. They are developing an understanding of functions and coordinate systems.

LEVELS OF SCIENCE PROFICIENCY

Level 150--Knows Everyday Science Facts

Students at this level know some general scientific facts of the type that could be learned from everyday experiences. They can read simple graphs, match the distinguishing characteristics of animals, and predict the operation of familiar apparatus that work according to mechanical principles.

Level 200--Understands Simple Scientific Principles

Students at this level are developing some understanding of simple scientific principles, particularly in the Life Sciences. For example, they exhibit some rudimentary knowledge of the structure and function of plants and animals.

Level 250--Applies Basic Scientific Information

Students at this level can interpret data from simple tables and make inferences about the outcomes of experimental procedures. They exhibit knowledge and understanding of the Life Sciences, including a familiarity with some aspects of animal behavior and of ecological relationships. These students also demonstrate some knowledge of basic information from the Physical Sciences.

Level 300--Analyzes Scientific Procedures and Data

Students at this level can evaluate the appropriateness of the design of an experiment. They have more detailed scientific knowledge, and the skill to apply their knowledge in interpreting information from text and graphs. These students also exhibit a growing understanding of principles from the Physical Sciences.

Level 350--Integrates Specialized Scientific Information

Students at this level can infer relationships and draw conclusions using detailed scientific knowledge from the Physical Sciences, particularly Chemistry. They also can apply basic principles of genetics and interpret the societal implications of research in this field.

LEVELS OF READING PROFICIENCY

Rudimentary (150)

Readers who have acquired rudimentary reading skills and strategies can follow brief written directions. They can also select words, phrases, or sentences to describe a simple picture and can interpret simple written clues to identify a common object. *Performance at this level suggests the ability to carry out simple, discrete reading tasks.*

Basic (200)

Readers who have learned basic comprehension skills and strategies can locate and identify facts from simple informational paragraphs, stories, and news articles. In addition, they can combine ideas and make inferences based on short, uncomplicated passages. *Performance at this level suggests the ability to understand specific or sequentially related information.*

Intermediate (250)

Readers with the ability to use intermediate skills and strategies can search for, locate, and organize the information they find in relatively lengthy passages and can recognize paraphrases of what they have read. They can also make inferences and reach generalizations about main ideas and author's purpose from passages dealing with literature, science, and social studies. *Performance at this level suggests the ability to search for specific information, interrelate ideas, and make generalizations.*

Adept (300)

Readers with adept reading comprehension skills and strategies can understand complicated literary and informational passages, including material about topics they study at school. They can also analyze and integrate less familiar material and provide reactions to and explanations of the text as a whole. *Performance at this level suggests the ability to find, understand, summarize, and explain relatively complicated information.*

Advanced (350)

Readers who use advanced reading skills and strategies can extend and restructure the ideas presented in specialized and complex texts. Examples include scientific materials, literary essays, historical documents, and materials similar to those found in professional and technical working environments. They are also able to understand the links between ideas even when those links are not explicitly stated and to make appropriate generalizations even when the texts lack clear introductions or explanations. *Performance at this level suggests the ability to synthesize and learn from specialized reading materials.*

LEVELS OF CIVICS PROFICIENCY

Level 200--Recognizes the Existence of Civic Life

Students at this level have a rudimentary knowledge of civics. They possess a beginning political awareness of the distinctions between the public and private domains and are familiar with some of the functions of government that pervade their immediate experience. They have some knowledge about elections and are developing an awareness of democratic principles such as the rule of law, as evidenced by their understanding that laws apply to government officials. These students also recognize that individuals--specifically the accused--have rights. Their elementary political vocabulary includes such terms as candidate, ballot, vice-president, judge, juror, and citizen.

Level 250--Understands the Nature of Political Institutions and the Relationship Between Citizen and Government

Students at this level are developing a knowledge of the nature of democratic institutions and processes. For example, they recognize the value of having more than one candidate in an election and the importance of the secret ballot. They are aware of the functions of a variety of government institutions and display a beginning understanding of federalism, as indicated by their ability to recognize the responsibilities of different levels of government. These students are developing an understanding of the reciprocal relationship between citizen and government. In addition to perceiving the purpose of individual rights in a democratic society and being able to identify some of these rights, such as the right to vote, they know of alternative ways to influence government--for example, making public speeches or writing letters to public officials. These students are developing a broader and more diverse political vocabulary.

Level 300--Understands Specific Government Structures and Functions

At this level, students have a more differentiated understanding of the structures, functions, and powers of American government as prescribed in the Constitution. For example, they have an increased understanding of federalism, are aware of the separation and allocation of powers, and grasp the concept of judicial review. These students are also familiar with certain historical events and legal precedents that have helped to shape our democratic heritage. They can apply their knowledge of individual rights to particular situations, and their conception of citizen action now includes cooperative political activity, such as boycotts and lobbying. These students are familiar with such terms as chief executive, constitutional rights, veto, and lobbyist, indicating an increasing understanding of the language of American politics. They can apply their civic knowledge to a large number and variety of complex situations.

Level 350--Understands a Variety of Political Institutions and Processes

Students at this level are distinguished by their broader and more detailed knowledge of the various institutions of government. For example, they can describe the responsibilities of the president, the Congressional power to override presidential vetoes and levy taxes, and the practice of judicial review. These students have a more elaborated understanding of a range of political processes--for example, presidential campaigns, primary elections, and public opinion polls. Their expanding political vocabulary includes such specialized terms as closed primary, impeachment, referendum, and recall election.

LEVELS OF U.S. HISTORY PROFICIENCY

Level 200--Knows Simple Historical Facts

Students at this level know some historical facts of the type learned from everyday experiences. For example, they can identify a few national holidays and patriotic symbols. They can read simple timelines, graphs, charts, and maps.

Level 250--Knows Beginning Historical Information and Has Rudimentary Interpretive Skills

Students at this level know a greater number and variety of historical facts of the type commonly learned from historical studies. For example, they can identify a number of historical figures, events, and terms. They are developing a sense of chronology and can interpret timelines, maps, and graphs.

Level 300--Understands Basic Historical Terms and Relationships

Students at this level have a broad knowledge of historical terms, facts, regions, and ideas. They have a general sense of chronology and can recognize characterizations of particular time periods in history. These students have some knowledge of the content of primary texts in U.S. political and constitutional history, such as the Declaration of Independence, Constitution, Bill of Rights, and Emancipation Proclamation. They are familiar with certain historically significant economic and social developments and have some awareness of different social and cultural groups. These students are beginning to comprehend the historical significance of domestic governmental policies and also the international context of U.S. history, as reflected in wars, exploration, settlements, immigration, and alliances. They show an emerging understanding of causal relationships.

Level 350--Interprets Historical Information and Ideas

Students at this level are developing a detailed understanding of historical vocabulary, facts, regions, and ideas. They are familiar with the content of a wider variety of texts, such as the Articles of Confederation, the Federalist Papers, Washington's Farewell Address, and certain amendments to the Constitution. They are aware of the religious diversity of the United States and recognize the continuing tension between democratic principles and such social realities as poverty and discrimination. These students demonstrate a rudimentary understanding of the history of U.S. foreign policy. They are beginning to relate social science concepts--such as price theory, separation of powers, and essential functions of government--to historical themes and can evaluate causal relationships.

#17. Question: What types of reports will NAEP produce based on the 1990 and 1992 national and state assessments, and when will they be available?

Answer: *In comparison to the shorter computerized score results produced by traditional standardized testing programs, NAEP publishes comprehensive and detailed reports on students' knowledge and understandings within each subject area assessed. In addition, the reports include information on the relationships between proficiency and an extensive range of background and instructional variables.*

More specifically, each state participating in the trial state assessment will receive multiple copies of an attractive 50-page summary (text and graphics) of its results, including national and regional comparisons. These reports will be produced by NAEP's new computerized report-generation system, in which graphic designers, data analysts, and report writers work to develop "shells" of the reports in advance of analysis. When the data are ready, they are automatically incorporated into the reports.

In addition, each state will receive complete documentation of its results in the form of four comprehensive data almanacs: (1) item-level results for the cognitive questions, showing the percentage of students in the state responding to each option for each question (almanac of item percents, with the correct response options marked with an asterisk), (2) the percentage of students responding to each category for the background questions and the level of subject-area proficiency associated with each category of response (proficiency almanac for student questionnaires), (3) as linked to their teachers, the percentage of students and proficiency levels for each response on the teacher questionnaire (proficiency almanac for teacher questionnaires), and (4) the percentage of students and proficiency levels for each response to each question on the school characteristics and policies questionnaire (proficiency almanac for school questionnaires). The results will be accompanied by a technical report documenting the procedures used in the trial state assessment.

Each state also will receive multiple copies of a 200-page composite or comparative report, including the same set of results for all participants as well as for the nation at three grade levels. It is anticipated that the composite report will begin with an executive summary or a highlight of the results that can be printed and distributed separately. An appendix, prepared by NCES, will contain state-level information on a number of indicators collected by other projects.

In total, each state will receive six reports--its summary report of results, four detailed data almanacs, a composite report of results across states and the nation, and a technical summary. In addition, states can arrange to receive documented data tapes of their own results.

For the national assessments, NAEP will produce the typical subject-matter "report cards," similar to those for the 1984, 1986, and 1988 assessments, that chronicle trends and report what students know and can do. For both the 1990 and 1992 assessments, these reports will be issued for mathematics, reading, science, and writing.

As discussed throughout this document, NAEP is a complex, constantly evolving project, and the trial state assessment program initiated in 1990 has added new challenges: managing the process of implementing 40 to 50 individual state assessments and providing for comparisons among them and to the nation. Although the process has been standardized to the greatest possible extent, each state has unique characteristics and, therefore, conducting this unprecedented task in a fair and valid way is demanding.

Thus, as shown below, the schedule indicates a much longer timeline for the state reports in 1990 than in 1992, particularly in scaling activities. However, for 1990, NAEP is responsible for developing scaling techniques that provide for fair comparisons across states, and of states to the nation. Compared to NAEP's last previous mathematics assessment in 1986, the grade levels are different, the content classifications for the subscales are different, the questions are different, the background variables are different, and the number of discrete datasets is very different with the addition of the

40 participating states and other entities. Finally, high visibility, congressionally mandated evaluation, and the potentially "high-stakes" nature of the trial state assessment program all suggest that 1990 is the time to increase quality-control efforts.

Schedule for 1990 and 1992 NAEP State Mathematics Assessments

	1990	1992
Complete Data Collection	May 25, 1990	March 22, 1992
Complete Scoring (4 to 6 million open-ended responses)	June 30, 1990	April 15, 1992
Complete Scaling	December 31, 1990	July 31, 1992
Complete Analysis	February 28, 1991	September 15, 1992
Camera-ready Copy for Reports (includes NAGB, state reviews)	May 31, 1991	December 15, 1992

Thus, for 1990, all reports pertaining to the mathematics assessments will be produced, reviewed, and in camera-ready form by May 31, 1991. This includes the reports and almanacs for each state, the composite report containing the national results and the results for the 40 participating states and entities, the technical report for the trial state program, and the trend report for the national trend assessment. The remaining trend reports will follow--a long-term reading trend report (1971 to 1990) for 9-, 13-, and 17-year-olds; a short-term reading trend report for fourth, eighth and twelfth graders (1988 to 1990); a long-term trend report for science (1969 to 1990) for 9-, 13- and 17-year-olds that reports the results for the 1990 assessment of fourth, eighth, and

twelfth graders on the same trend scale; and a writing trend report (1984 to 1990) for fourth, eighth, and eleventh graders that is linked to the results of the 1990 portfolio study at grades 4 and 8.

For 1992, all the reports pertaining to the trial state assessment in mathematics at grades 4 and 8 will be camera-ready by December 15, 1992, or nearly six months earlier than scheduled for the first phase of the trial state assessment program. Again, this includes the reports and almanacs for each state, the composite reports containing the comparative national results and results for the participating states, and a technical report documenting the procedures used in the 1992 trial state assessment program. The complete set of state and national reports for the all-new 1992 reading assessment and the reports containing the national information about trends and what students know and can do in reading, writing, mathematics, and science will follow, all in camera-ready form, no later than May 31, 1993. The state reading reports will require extensive development and review, as this will be the first time this information is collected, analyzed, and reported. The trend reports must await the end of data collection in late May, and also take some additional time to produce because they are authored rather than computer-generated.

In addition to the assessment reports, NAEP also publishes a detailed technical report and creates a well-documented public-use data tape for each assessment. These data tapes can be used by analysts external to NAEP to conduct in-depth studies using NAEP data.

FURTHER READING

Additional information about NAEP procedures and results can be found in the following selected publications:

Russell Allen, Norman Bettis, Walter B. MacDonald, Ina V.S. Mullis, and Christopher Salter, The Geography Learning of High-School Seniors (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1990).

Arthur N. Applebee, Judith A. Langer, Ina V.S. Mullis, and Lynn B. Jenkins, The Writing Report Card, 1984-88 (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1990).

Arthur N. Applebee, Judith A. Langer, Ina V.S. Mullis, Lynn B. Jenkins, and Mary A. Foertsch, Learning to Write in Our Nation's Schools (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, in press).

Lee Anderson, Lynn B. Jenkins, James Leming, Walter B. MacDonald, Ina V.S. Mullis, Mary Jane Turner, and Judith Wooster, The Civics Report Card (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1990).

Albert Beaton, Implementing the New Design: The NAEP 1983-84 Technical Report (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1987).

Albert Beaton, Expanding the New Design: The NAEP 1985-86 Technical Report (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1988).

John A. Dossey, Ina V.S. Mullis, Mary M. Lindquist, Donald L. Chambers, The Mathematics Report Card: Are We Measuring Up? (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1988).

David Hammack, Michael Hartoonian, John Howe, Lynn B. Jenkins, Linda S. Levstik, Walter B. MacDonald, Ina V.S. Mullis, and Eugene Owen, The U.S. History Report Card (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1990).

Educational Testing Service, Mathematics Objectives, 1990 Assessment (Princeton, NJ: Educational Testing Service, 1988).

Educational Testing Service, Science Objectives, 1990 Assessment (Princeton, NJ: Educational Testing Service, 1989).

Eugene G. Johnson and Rebecca Zwick, The NAEP 1988 Technical Report (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, in press).

Judith A. Langer, Arthur N. Applebee, Ina V.S. Mullis, and Mary A. Foertsch, Learning to Read in Our Nation's Schools (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, in press).

Ina V.S. Mullis and Lynn B. Jenkins, The Science Report Card: Elements of Risk and Recovery (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1988).

Ina V.S. Mullis and Lynn B. Jenkins, The Reading Report Card, 1971-88 (Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress, 1990).

66